ASSIGNMENT – 2 MACHINE LEARNING

Q1 to Q11 have only one correct answer. Choose the correct option to answer your question.

1.Movie Recommendation systems are an example of:

i) Classification

ii) Clustering

iii) Regression Options: a) 2 Only

b) 1 and 2

c) 1 and 3

d) 2 and 3

Ans: d)  
  
Generally, movie recommendation systems cluster the users in a finite number of similar groups based on their previous activities and profile. Then, at a fundamental level, people in the same cluster are made similar recommendations.  
  
In some scenarios, this can also be approached as a classification problem for assigning the most appropriate movie class to the user of a specific group of users. Also, a movie recommendation system can be viewed as a reinforcement learning problem where it learns by its previous recommendations and improves the future recommendations.

2. Sentiment Analysis is an example of:

i) Regression

ii) Classification

iii) Clustering

iv) Reinforcement Options:

a) 1 Only

b) 1 and 2

c) 1 and 3

d) 1, 2 and 4

Ans: d)

Sentiment analysis at the fundamental level is the task of classifying the sentiments represented in an image, text or speech into a set of defined sentiment classes like happy, sad, excited, positive, negative, etc. It can also be viewed as a regression problem for assigning a sentiment score of say 1 to 10 for a corresponding image, text or speech.  
  
Another way of looking at sentiment analysis is to consider it using a reinforcement learning perspective where the algorithm constantly learns from the accuracy of past sentiment analysis performed to improve the future performance.

3. Can decision trees be used for performing clustering?

a) True

b) False

Ans: a)

Decision trees can also be used to for clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

i) Capping and flooring of variables

ii) Removal of outliers Options:

a) 1 only

b) 2 only

c) 1 and 2

d) None of the above

Ans: a)

Removal of outliers is not recommended if the data points are few in number. In this scenario, capping and flouring of variables is the most appropriate strategy.

5. What is the minimum no. of variables/ features required to perform clustering?

a) 0

b) 1

c) 2

d) 3

Ans: b)

At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram.

6. For two runs of K-Mean clustering is it expected to get same clustering results?

a) Yes

b) No

Ans: b)

K-Means clustering algorithm instead converses on local minima which might also correspond to the global minima in some cases but not always. Therefore, it's advised to run the K-Means algorithm multiple times before drawing inferences about the clusters.  
  
However, note that it's possible to receive same clustering results from K-means by setting the same seed value for each run. But that is done by simply making the algorithm choose the set of same random no. for each run.

7. Is it possible that Assignment of observations to clusters does not change between successive

iterations in K-Means?

a) Yes

b) No

c) Can't say

d) None of these

Ans: a)

When the K-Means algorithm has reached the local or global minima, it will not alter the assignment of data points to clusters for two successive iterations.

8. Which of the following can act as possible termination conditions in K-Means?

i) For a fixed number of iterations.

ii) Assignment of observations to clusters does not change between iterations. Except for cases witha bad local minimum.

iii) Centroids do not change between successive iterations.

iv) Terminate when RSS falls below a threshold.

Options:

a) 1, 3 and 4

b) 1, 2 and 3

c) 1, 2 and 4

d) All of the above

Ans: d)

All four conditions can be used as possible termination condition in K-Means clustering:  
  
This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.  
Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.  
This also ensures that the algorithm has converged at the minima.  
Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it's a good practice to combine it with a bound on the number of iterations to guarantee termination.

9. Which of the following algorithms is most sensitive to outliers?

a) K-means clustering algorithm

b) K-medians clustering algorithm

c) K-modes clustering algorithm

d) K-medoids clustering algorithm

Ans: a)

Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

i) Creating different models for different cluster groups.

ii) Creating an input feature for cluster ids as an ordinal variable.

iii) Creating an input feature for cluster centroids as a continuous variable.

iv) Creating an input feature for cluster size as a continuous variable. Options:

a) 1 only

b) 2 only

c) 3 and 4

d) All of the above

Ans: a)

Creating an input feature for cluster ids as ordinal variable or creating an input feature for cluster centroids as a continuous variable might not convey any relevant information to the regression model for multidimensional data. But for clustering in a single dimension, all of the given methods are expected to convey meaningful information to the regression model. For example, to cluster people in two groups based on their hair length, storing clustering ID as ordinal variable and cluster centroids as continuous variables will convey meaningful information.

11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

a) Proximity function used

b) of data points used

c) of variables used

d) All of the above

Ans: d)

Change in either of Proximity function, no. of data points or no. of variables will lead to different clustering results and hence different dendrograms.

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly

12. Is K sensitive to outliers?

[**K-Means clustering**](https://medium.com/@joel_34096/k-means-clustering-using-python-from-scratch-7ccdace7789) is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.

But sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the rest of data points. Let us look at one method for finding outliers of univariate data (one dimensional).

The lower quartile ‘Q1’ is median of first half of data. The upper quartile ‘Q3’ is median of second half of data. The interquartile range ‘IQR’ is difference of Q3 and Q1. An outlier is a point that is greater than (Q3 + 1.5\*IQR) or lesser than (Q1–1.5\*IQR). The given below code can be used to find the outliers.

Q1 = np.percentile(data, 25, interpolat**ion = ‘midpoint’)**

# The lower quartile Q1 is calculated.**Q3 = np.percentile(data, 75, interpolation = ‘midpoint’)**

# The upper quartile Q3 is calculated.**IQR = Q3 — Q1**

# The Interquartile range is calculated.**Q3 + 1.5\*IQR, Q1–1.5\*IQR**

# The outlier range is calculated.

Let us take an example to understand how outliers affect the mean of data using python.

**X = list(np.random.rand(100))** # ‘X’ is a list of 100 random numbers between 0 and 1.  
**Y = list(np.linspace(1,10,100))** # ‘Y’ is a list of 100 random numbers equally spaced between 1 and 10.**plt.figure(figsize=(20,10))** # Size of figure is adjusted.  
**plt.xticks(fontsize=20)** # Size of number labels on x-axis is adjusted.  
**plt.yticks(fontsize=20)** # Size of number labels on y-axis is adjusted.  
**plt.xlabel(‘X Values’,fontsize=20)** # x-axis is labelled.  
**plt.ylabel(‘Y Values’,fontsize=20)** # y-axis is labelled.**mean\_X = sum(X)/len(X)** # ‘mean\_X’ is the mean value of ‘X’.  
**mean\_Y = sum(Y)/len(Y)** # ‘mean\_Y’ is the mean value of ‘Y’.  
**plt.plot(mean\_X,mean\_Y,’ro’,markersize = 10)** # The mean value (mean\_X,mean\_Y) point is plotted.**outlier = 1000** # An outlier of value 1000.  
**X.append(outlier)** # The outlier is added to ‘X’.  
**Y.append(Y[99] + Y[1] — Y[0])** # An extra number is added to ‘Y’ such equal spacing still holds.**mean\_X\_new = sum(X)/len(X)** # ‘mean\_X\_new’ is new mean value of ‘X’.  
**mean\_Y\_new = sum(Z)/len(Z)** # ‘mean\_Y\_new’ is new mean value of ‘Y’.  
**plt.plot(mean\_X\_new,mean\_Y\_new,’go’,markersize = 10)** # The mean value (mean\_X,mean\_Y) point is plotted in green.

Chart

Description automatically generated with medium confidence

The red point is mean of data excluding outlier. The green point is mean of data including outlier.

We observe that the outlier increases the mean of data by about 10 units. This is a significant increase considering the fact that all data points range from 0 to 1. This shows that the mean is influenced by outliers.

Since K-Means algorithm is about finding mean of clusters, the algorithm is influenced by outliers. Let us take an example to understand how outliers affect the K-Means algorithm using python.

We have a 2 dimensional data set called ‘cluster’ consisting of 3000 points with no outliers. We get the following scatter plot after K-means algorithm is applied.

Chart, scatter chart

Description automatically generated

Now we add 60 outliers to ‘cluster’ data set. The outliers is about 2 percent of non-outliers. We get the following scatter plots for different values of outliers after K-means algorithm is applied.

Chart, scatter chart

Description automatically generated

The outliers are not shown in the scatter plot. Only the 3000 non outlier points is shown in the scatter plot for sake of better visualisation. The outliers form a seperate cluster represented by centroid number = 3.

Chart, scatter chart

Description automatically generated

The outliers are not shown in the scatter plot. Only the 3000 non outlier points is shown in the scatter plot for sake of better visualisation. The outliers form a seperate cluster represented by centroid number = 3.

We observe that the outliers show up as a separate cluster and also cause other clusters to merge which suggests clustering was not efficient when outliers were included in data set.

Even though the outliers were about 2 percent of non-outliers which is common in real world data sets, they had a significant impact on clustering. Hence it is better to identify and remove outliers before applying K-means clustering algorithm.

13. Why is K means better?

One of the major advantages of K-Means is that it can handle larger data sets compared to for example the hierarchical cluster approaches. However, K-Means comes with some disadvantages as well. It is sensitive to outliers.

14. Is K means a deterministic algorithm?

A deterministic function always returns the same results if given the same input values. ... A nondeterministic function may return different results every time it is called, even when the same input values are provided.

The basic **k-means** clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids. ... Moreover, it provides comparatively better predictions of cancer subtypes from gene expression data./span> . However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method.